DOCUMENT RESUME

ED 300 446 TM 012 495

AUTHOR Miller, Sherri K.; And Others
TITLE Differential Item Performance for Mexican-American ESL Students and White Non-ESL Students on Mathematics and English Achievement Tests.
PUB DATE Apr 88
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; College Entrance Examinations; English; *English (Second Language); High Schools; High School Students; Item Analysis; Language Proficiency; Language Tests; Mathematics Achievement; *Mathematics Tests; *Mexican Americans; Test Bias; *White Students
IDENTIFIERS *ACT Assessment; *Differential Item Performance

ABSTRACT

In the fall of 1986, a pilot study was conducted to investigate the differential performance at the item level of Mexican-American students who spoke English as a second language (ESL) versus White native English speakers. This study was designed to replicate the pilot study and test the hypotheses based on that study. The test materials used were the English Usage and Mathematics Usage tests of the American College Testing Program Assessment. It was hypothesized that: (1) items that emphasize mechanics in the English Usage Test, such as grammar and punctuation, tend to favor ESL examinees; (2) items that focus upon style and structure in the English Usage Test tend to favor non-ESL students; and (3) mathematical items with the greatest verbal load tend to favor non-ESL examinees. Subjects included 471 Mexican-American ESL students and 1,000 White native speakers (non-ESL). Results indicate that none of the three hypotheses was supported. Although the mean score for the ESL students was almost a full standard deviation below that of the native speakers for both tests, it appears that the group difference in performance was reflected throughout most of the test items. Specific categories of items that were disproportionately easy or hard for either of the groups could not be found. Six graphs and four tables conclude the document. (TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED300446

# DIFFERENTIAL ITEM PERFORMANCE FOR MEXICAN-AMERICAN ESL STUDENTS AND WHITE NON-ESL STUDENTS ON MATHEMATICS AND ENGLISH ACHIEVEMENT TESTS

Sherri K. Miller
Allen E. Doolittle
Terry A. Ackerman

Presented at the 1988 NCME Annual Meeting,
April 5, New Orleans, LA

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SHERRI K. MILLER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012 495

# DIFFERENTIAL ITEM PERFORMANCE FOR MEXICAN-AMERICAN ESL STUDENTS AND WHITE NON-ESL STUDENTS ON MATHEMATICS AND ENGLISH ACHIEVEMENT TESTS

## Introduction

Test bias, whether associated with race, sex, or other population subgroups, is a serious and highly complex issue. Test and item bias are conceptualized as something that invalidates the meaning of test results for some subgroup of the population. One subgroup of particular interest is "English as a second language" (ESL) students. Performance on admissions and placement tests may be affected in nontrivial ways by the English language proficiency of ESL students.

ESL students represent many different ethnic groups. The culture and language of the different ethnic groups vary widely; therefore, it is preferable to study ESL students separately for each ethnic group. The present study examines the performance of Mexican-American ESL students.

Several researchers have looked at the validity of standardized tests for Mexican-American ESL students in the context of an external criterion, viz., college performance (Alderman, 1982; Breland and Duran, 1985; Mestre, 1981). The results of these studies vary. Some found evidence of differential validity while others did not. Differences in such things as predictor variables, criterion variables, and sample composition make comparisons among these studies difficult.

It is possible to think of validity in the absence of an external criterion. Internal analysis focuses on the group performance within a measure. In the fall of 1986, a pilot study was conducted that investigated the differential performance of Mexican-American ESL students and white non-ESL students at the item level. The test materials used in the pilot study were the English Usage and Mathematics Usage tests of the ACT Assessment

administered in October 1985. A procedure developed by Mantel and Haenszel (1959) was used to examine differential item performance.

For the English Usage Test, 10 of the 75 items were identified as performing significantly different for ESL and non-ESL examinees. ESL students were favored on five of the 10 items. When looking at the content classification of each of these ten items, there did not appear to be any systematic differences in the classification of items that favored ESL and non-ESL examinees. However, when examining all items in the test, both significant and nonsignificant, the white non-ESL students tended to perform relatively better on logic and organization items, while the ESL students tended to perform relatively better on grammar items.

For the Mathematics Usage Test, none of the items showed a significant difference between the Mexican-American ESL and white non-ESL examinees. In looking at the direction of the Mantel-Haenszal statistic for the different categories of math items, there again did not appear to be a systematic difference between ESL examinees and white non-ESL examinees. However, there seemed to be a slight tendency for the number of words in story problems to be related to the degree that the items seemed to favor the non-ESL examinees.

The present study was designed to replicate the pilot study and test the hypotheses formed on the basis of that study. The hypotheses formed are listed below:

1. Items that emphasize mechanics in the English Usage Test, such as grammar and punctuation, tend to favor ESL examinees.
2. Items that focus upon style and structure in the English Usage Test tend to favor non-ESL students.
3. Mathematical items with the greatest verbal load tend to favor non-ESL examinees.

The primary objective of the present study was to investigate English usage and mathematics items for differential item performance based on ESL and non-ESL examinees. A second objective of the study was to investigate specific hypotheses about the items with respect to differential item performance.

## Methodology

### Instrument and Subjects

The test materials used in the present study were the English Usage and Mathematics Usage tests of the ACT Assessment, a college entrance exam. The English Usage Test is a 75-item, 40-minute test that measures understanding and use of basic elements of correct and effective writing: punctuation, grammar, sentence structure, diction and style, and logic and organization. The Mathematics Usage Test is a 40-item, 50-minute test that measures mathematical reasoning ability in six content areas. See Table 1 for a list and description of the item categories in each test.

The samples of 471 Mexican-American, self-reported ESL students and 1000 white self-reported non-ESL students were taken from the October 1986 ACT Assessment administration. All Mexican-American ESL students who took the ACT Assessment in October 1986 were included in the study. The 1000 white non-ESL students were randomly selected from the group of 160,220 white non-ESL examinees who took the ACT Assessment on that same date.

- - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - -

Index of Differential Item Performance

A contingency table procedure was used to measure differential item performance (Mantel & Haenszel, 1959). The Mantel-Haenszel statistic (MH-CHISQR, see Holland and Thayer, 1986) is based upon 2 x 2 contingency tables for each total score category. The MH-CHISQR statistic is distributed as a chi-square with one degree of freedom and is therefore a powerful unbiased test (Cox, 1970). Two statistics related to the MH-CHISQR, $\hat{\alpha}_{MH}$ and $\hat{z}_{MH}$, were also examined. The common odds ratio, $\hat{\alpha}_{MH}$, across the 2 x 2 tables, is given by

$$\hat{\alpha}_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j} .$$

Where $T_j$ is the total number of examinees in the jth matched set. $A_j$ and $C_j$ represent the number of examinees in the reference and focal groups who answered an item correctly. $B_j$ and $D_j$ are the number of examinees who responded incorrectly from the reference and focal groups. The reference group establishes a standard against which the performance of the focal group is compared. This ratio is on a scale of 0 to $\infty$ with $\alpha = 1$ representing a null value or no differential item performance.

The value of $\hat{\alpha}_{MH}$, for a studied item, is the "average factor by which the odds that a member of the reference group is correct on the studied item exceeds the corresponding odds for a comparable member of the focal group"

(Holland and Thayer, 1986). Holland and Thayer suggest taking the log of $\hat{\alpha}_{MH}$ to put it into a symmetric scale with zero as the null value. We propose a slight modification of this procedure,

$$\hat{z}_{MH} = -\frac{1}{1.7} \ln (\hat{\alpha}_{MH}),$$

as a measure of the amount of differential item performance.

The value of $\hat{z}_{MH}$ is a measure of the degree to which a white non-ESL examinee found the studied item more difficult than did a comparably-scoring ESL examinee. Positive values imply that the ESL examinees found the item relatively easier than the white non-ESL examinees; negative values indicate that ESL examinees found the item relatively harder.

## Results

Means, standard deviations, and sample size for each group are presented in Table 2. Mean raw scores and standard deviations are lower for Mexican-American ESL examinees than for white non-ESL examinees for both tests.

- - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - -

Figures 1 and 2 present the cumulative frequency polygons for the Mexican-American ESL sample and the white non-ESL sample for the English Usage Test and the Mathematics Usage Test, respectively.

- - - - - - - - - - - - - - - - - - - - -

Insert Figures 1 and 2 about here

- - - - - - - - - - - - - - - - - - - - -

In Table 3, the $z$ statistics from the M-H analysis, comparing Mexican-American ESL examinees and white non-ESL examinees, are reported for the mathematics and English items. A baseline for judging the magnitude of the $z_{MH}$ statistic was obtained from an analysis which compared two randomly equivalent groups of 1,000 white non-ESL examinees (see Shepard, 1984). Index values that exceeded the largest value occurring in the white-white analysis (.20 for English Usage and .23 for Mathematics) are starred in Table 3 as performing differentially.

- - - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - - -

A substantial number of items at the end of the English test were flagged as easier for white non-ESL examinees as compared to Mexican-American ESL examinees. Figure 3 displays the magnitude and direction of the $z_{MH}$ statistic pictorially with the 75 items grouped according to passage set. As can be seen, the last two passages contain more items which favor white non-ESL examinees. We speculated that this was due to a differential speededness effect between white non-ESL and Mexican-American ESL examinees. Since items omitted by examinees do not enter into our computation of the M-H statistics, we further speculated that the speededness effect was showing up because Mexican-American ESL examinees, in an effort to finish the test, may have randomly answered the last items. In an effort to reduce this possible differential speededness effect on the English Test, the M-H analysis was again done on the English Test, this time with the last two passage sets omitted, and thereby reducing the number of items from 75 to 58. Table 4 shows the items flagged in the 58-item English Test using the index value obtained in the white-white comparison as the criterion. (The criterion index

value was unchanged since there was no evidence of a speededness effect for the white non-ESL examinees.)

- - - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - - -

Results for the English Usage Test

As can be seen in Table 4, seven items were flagged as performing differentially. Three of the items were found to be relatively easier for Mexican-American ESL examinees and four were found to be relatively easier for the white non-ESL examinees. Each of the three items found to favor the Mexican-American ESL examinees were classified as diction and style items. For the four items found to favor the white non-ESL examinees, one was classified as a punctuation item, two were sentence structure items, and one a diction and style item. Looking only at the seven items flagged as performing differentially, no conclusive evidence of any systematic differences in the classification of items that favored ESL and non-ESL examinees could be found. There was, at most, only a hint of a tendency for sentence structure items to favor non-ESL examinees and for diction and style items and grammar items to favor ESL examinees (see Figure 4).

- - - - - - - - - - - - - - -

Insert Figure 4 about here

- - - - - - - - - - - - - - -

Results for the Mathematics Usage Test

For the Mathematics Usage Test, only two items were flagged as performing differentially (see Table 3). Both of these items were arithmetic and algebraic reasoning items and both favored white non-ESL examinees. When all items in the test were examined, both significant and nonsignificant, there appeared to be no systematic differences in the content classification of items favoring ESL examinees and white non-ESL examinees (see Figure 5).

Items were also classified according to their verbal load. Figure 6 shows the magnitude and direction of the M-H Z statistic for all items categorized as either, (1) equations only (no words), (2) standard word count[1] less than 40, and (3) standard word count greater than or equal to 40. In general, the hypothesis that high word-count items favor non-ESL students was not supported. However, the two items with the largest index values were high word-count items that did favor the non-ESL students.

- - - - - - - - - - - - - - - - - - -

Insert Figures 5 and 6 about here

- - - - - - - - - - - - - - - - - - -

## Discussion and Conclusions

The purpose of this research was to replicate the pilot study and test the hypotheses formed on the basis of the pilot study. Hypothesis 1, which stated that items emphasizing mechanics (such as grammar and punctuation) in

---

[1] standard word count here is defined as the number of characters in an item stem divided by 6.

the English Usage Test tend to favor ESL examinees, was not supported. Although more grammar items tended to favor ESL examinees than non-ESL examinees, the magnitude of the $z_{MH}$ statistic was less than .15 for each item. Also, items in the punctuation classification had a slight tendency to favor non-ESL examinees--the opposite of what was hypothesized.

Hypothesis 2, which stated that items that focus upon style and structure in the English Usage Test tend to favor non-ESL students, was not supported by the present research. Although sentence structure items seemed to favor non-ESL examinees, items classified as diction and style seemed to favor ESL examinees as seen in Figure 4, and items classified as logic and organization did not seem to favor either group.

Hypothesis 3, which stated that the verbal load of the math items is related to differential item performance for ESL and non-ESL examinees, was also not strongly supported. However, the two items that were flagged as favoring white non-ESL examinees were items with high word counts.

In summary, the results do not provide support for the specific hypotheses that were the focus of this study. Although the mean score for the Mexican-American ESL students was almost a full standard deviation below that of the non-ESL students for both tests, it appears that the group difference in performance was reflected throughout most of the items in the test. Both tests seemed to be functioning comparably for each of the investigated groups of examinees. We were unable to find specific categories of items that were disproportionately easy or hard for either of the groups.

## References

Alderman, D. (1982). Language proficiency as a moderator variable in testing academic aptitude. Journal of Educational Psychology, 74, 580-587.

Breland, H.M. and Duran, R.P. (1985). Assessing English composition skills in Spanish-speaking populations. Educational and Psychological Measurement, 45, 309-317.

Holland, P.W., & Thayer, D.T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studeies of disease. Journal of the National Cancer Institute, 22, 719-748.

Mestre, J.P. (1981). Predicting academic achievement among bilingual Hispanic college technical students. Educational and Psychological Measurement, 41, 1255-1264.

Shepard, L., Camilli, G., and Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Table 1

Content Categories of the ACT Assessment
English Usage and Mathematics Tests

## English Usage

**Punctuation.** The items in this category test such conventions as the use and placement of commas, colons, semicolons, dashes, parentheses, apostrophes, and quotation, question, and exclamation marks.

**Grammar.** The items in this category test adjectives and adverbs, conjuctions, and agreement between subject and verb and between pronouns and their antecedents.

**Sentence Structure.** The items in this category test relationships between/among clauses, placement of modifiers, parallelisms, and shifts in construction.

**Diction and Style.** The items in this category test precision in word choice, appropriateness in figurative language, and economy in writing.

**Logic and Organization.** The items in this category test the logical organization of ideas: paragraphing, transitions, unity, and coherence.

## Mathematics

**Arithmetic and Algebraic Operations.** The items in this category explicitly describe operations to be performed by the student. The operations include manipulating and simplifying expressions containing arithmetic or algebraic fractions, performing basic operations in polynomials, solving linear equations in one unknown, and performing operations on signed numbers.

**Arithmetic and Algebraic Reasoning.** These word problems present practical situations in which algebraic and/or arithmetic reasoning is required. The problems require the student to interpret the question and either to solve the problem or to find an approach to its solution.

**Geometry.** The items in this category cover such topics as measurement of lines and plane surfaces, properties of polygons, the Pythagorean theorem, and relationships involving circles. Both formal and applied problems are included.

**Intermediate Algebra.** The items in this category cover such topics as dependence and variation of quantities related by specific formulas, arithmetic and geometric series, simultaneous equations, inequalities, exponents, radicals, graphs of equations, and quandratic equations.

**Number and Numeration Concepts.** The items in this category cover such topics as rational and irrational numbers, set properties and operations, scientific notation, prime and composite numbers, numeration systems with bases other than 10, and absolute value.

**Advanced Topics.** the items in this category cover such topics as trigonometric functions, permutations and combinations, probability, statistics, and logic. Only simple applications of the skills implied by these topics are tested.

**Table 2**

Means, standard deviations, and sample size
for ESL and non-ESL examinees
for Mathematics and English Usage Tests

| | $\bar{X}_{\text{raw score}}$ | SD | N |
|---|---|---|---|
| English Usage (58 items) | | | |
| Mex.-Am. ESL | 31.50 | 9.40 | 471 |
| White non-ESL | 39.83 | 9.58 | 1000 |
| Mathematics (40 items) | | | |
| Mex.-Am. ESL | 14.90 | 7.54 | 471 |
| White non-ESL | 20.79 | 8.64 | 1000 |

Table 3

Mantel-Haenszel Z Index[1] for Mathematics and English Usage Items

| Item | English Usage (75 items) Mex-Am ESL vs White Non-ESL | Mathematics Usage (40 items) Mex-Am ESL vs White Non-ESL |
|---|---|---|
| 1 | -0.02 | -0.03 |
| 2 | -0.16 | -0.33* |
| 3 | -0.17 | -0.03 |
| 4 | 0.31* | 0.17 |
| 5 | 0.04 | 0.05 |
| 6 | -0.07 | 0.05 |
| 7 | 0.16 | 0.02 |
| 8 | 0.06 | -0.1. |
| 9 | -0.11 | 0.17 |
| 10 | 0.37* | 0.14 |
| 11 | 0.11 | -0.04 |
| 12 | -0.01 | 0.11 |
| 13 | 0.17 | 0.03 |
| 14 | 0.23* | 0.07 |
| 15 | -0.14 | -0.08 |
| 16 | -0.01 | -0.19 |
| 17 | -0.15 | 0.05 |
| 18 | 0.01 | -0.07 |
| 19 | 0.17 | 0.02 |
| 20 | -0.20* | -0.09 |
| 21 | -0.21* | 0.05 |
| 22 | 0.09 | 0.00 |
| 23 | 0.09 | 0.14 |
| 24 | 0.13 | 0.03 |
| 25 | 0.22* | -0.06 |
| 26 | 0.15 | 0.17 |
| 27 | 0.09 | 0.09 |
| 28 | 0.09 | -0.03 |
| 29 | -0.14 | -0.10 |
| 30 | 0.00 | -0.10 |
| 31 | 0.16 | -0.01 |
| 32 | 0.13 | 0.09 |
| 33 | 0.07 | 0.06 |
| 34 | 0.24* | 0.12 |
| 35 | 0.24* | 0.16 |
| 36 | -0.01 | -0.02 |
| 37 | -0.15 | -0.38* |
| 38 | 0.04 | -0.12 |
| 39 | 0.19 | 0.12 |
| 40 | -0.08 | -0.10 |
| 41 | -0.14 | |
| 42 | 0.00 | |
| 43 | 0.15 | |
| 44 | -0.11 | |
| 45 | -0.11 | |
| 46 | 0.08 | |
| 47 | 0.00 | |
| 48 | 0.22* | |
| 49 | 0.20* | |
| 50 | 0.22* | |
| 51 | 0.27* | |
| 52 | -0.02 | |
| 53 | -0.04 | |
| 54 | 0.11 | |
| 55 | 0.14 | |
| 56 | 0.17 | |
| 57 | 0.17 | |
| 58 | -0.35* | |
| 59 | -0.41 | |
| 60 | 0.08 | |
| 61 | 0.17 | |
| 62 | -0.35* | |
| 63 | 0.19 | |
| 64 | -0.19 | |
| 65 | -0.25* | |
| 66 | 0.11 | |
| 67 | -0.23* | |
| 68 | 0.12 | |
| 69 | -0.21* | |
| 70 | -0.05 | |
| 71 | 0.11 | |
| 72 | -0.30* | |
| 73 | 0.01 | |
| 74 | -0.12 | |
| 75 | -0.28* | |

[1] Negative values correspond to items that the Non-ESL group found easier on the average than did comparible ESL group members.

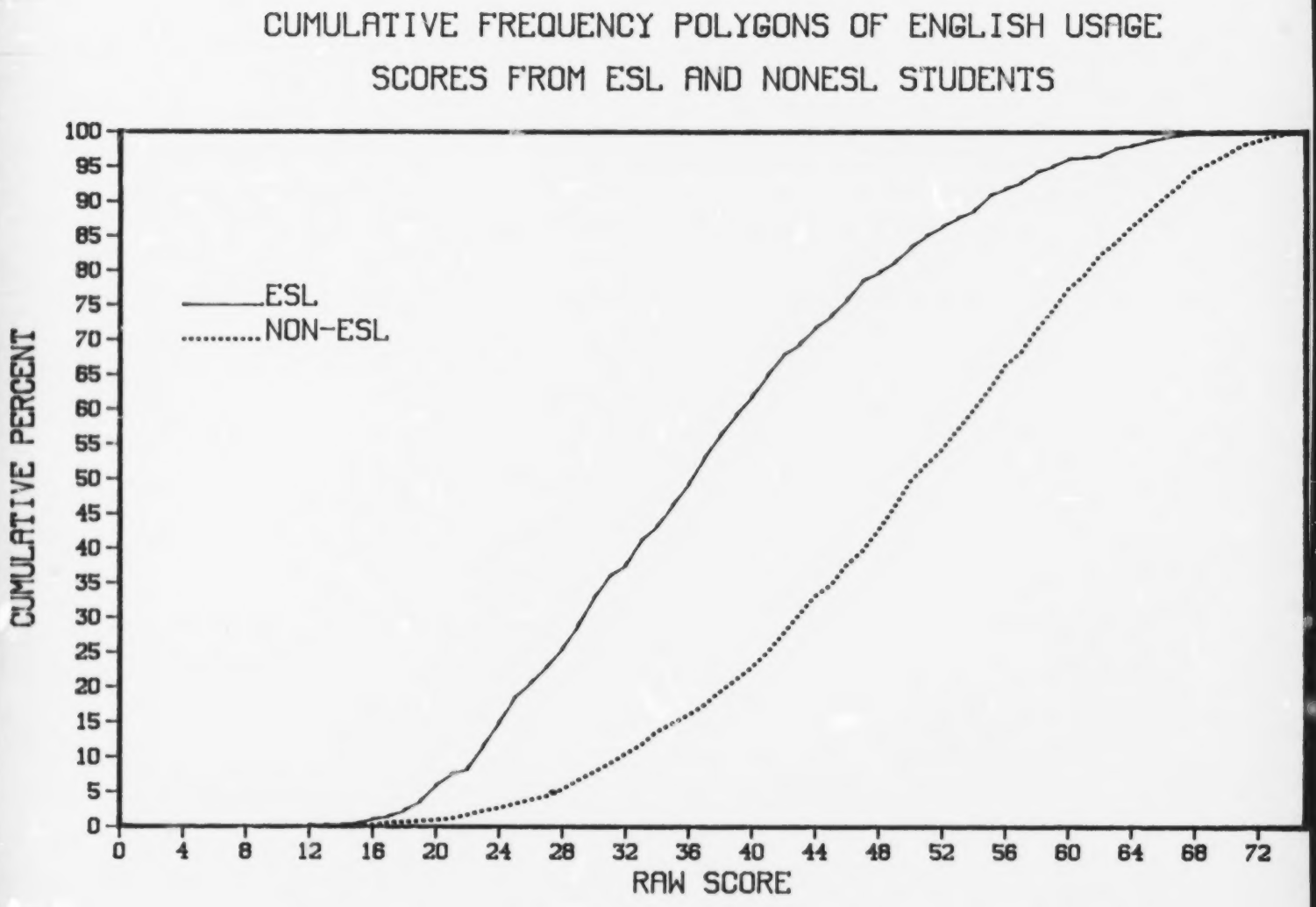
Table 4

Mantel-Haenszel Z Index[1] for the English Usage Test Items deleting the last two passage sets

| Item | Mex-Am ESL vs White Non-ESL |
|---|---|
| 1 | -0.05 |
| 2 | -0.18 |
| 3 | -0.18 |
| 4 | 0.31* |
| 5 | 0.00 |
| 6 | -0.13 |
| 7 | 0.17 |
| 8 | 0.01 |
| 9 | -0.12 |
| 10 | 0.33* |
| 11 | 0.09 |
| 12 | -0.11 |
| 13 | 0.14 |
| 14 | 0.17 |
| 15 | -0.13 |
| 16 | -0.07 |
| 17 | -0.19 |
| 18 | 0.00 |
| 19 | 0.11 |
| 20 | -0.25* |
| 21 | -0.25* |
| 22 | 0.04 |
| 23 | 0.07 |
| 24 | 0.10 |
| 25 | 0.15 |
| 26 | 0.11 |
| 27 | 0.05 |
| 28 | 0.05 |
| 29 | -0.15 |
| 30 | -0.02 |
| 31 | 0.12 |
| 32 | 0.10 |
| 33 | 0.04 |
| 34 | 0.18 |
| 35 | 0.21* |
| 36 | -0.05 |
| 37 | -0.19 |
| 38 | 0.04 |
| 39 | 0.19 |
| 40 | -0.14 |
| 41 | -0.25* |
| 42 | -0.03 |
| 43 | 0.07 |
| 44 | 0.06 |
| 45 | -0.17 |
| 46 | 0.08 |
| 47 | -0.10 |
| 48 | 0.18 |
| 49 | 0.13 |
| 50 | -0.02 |
| 51 | 0.15 |
| 52 | 0.19 |
| 53 | -0.08 |
| 54 | -0.05 |
| 55 | 0.09 |
| 56 | 0.08 |
| 57 | 0.11 |
| 58 | -0.37* |

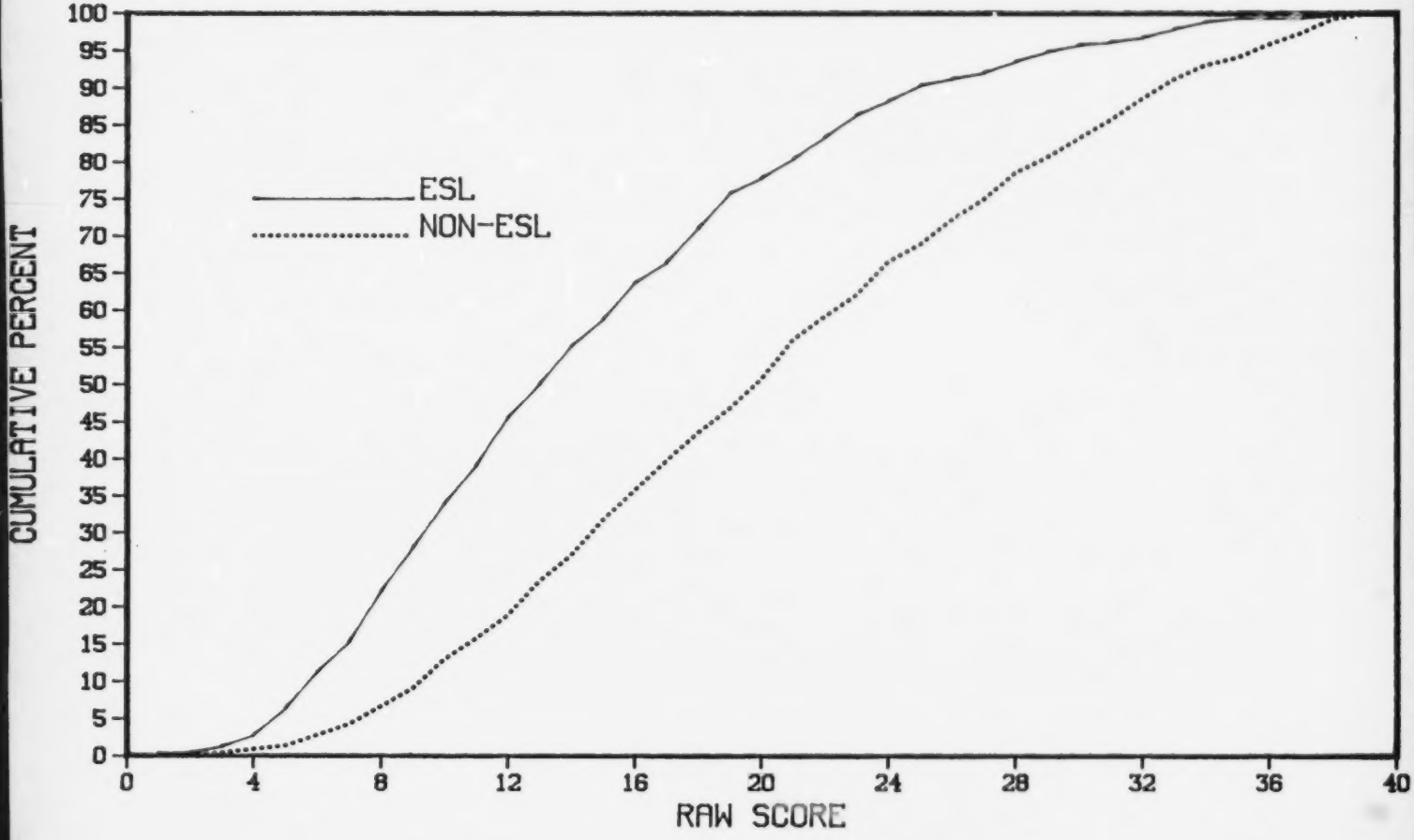
[1] Negative values correspond to items that the Non-ESL group found easier on the average than did comparible ESL group members.

# FIGURE 1

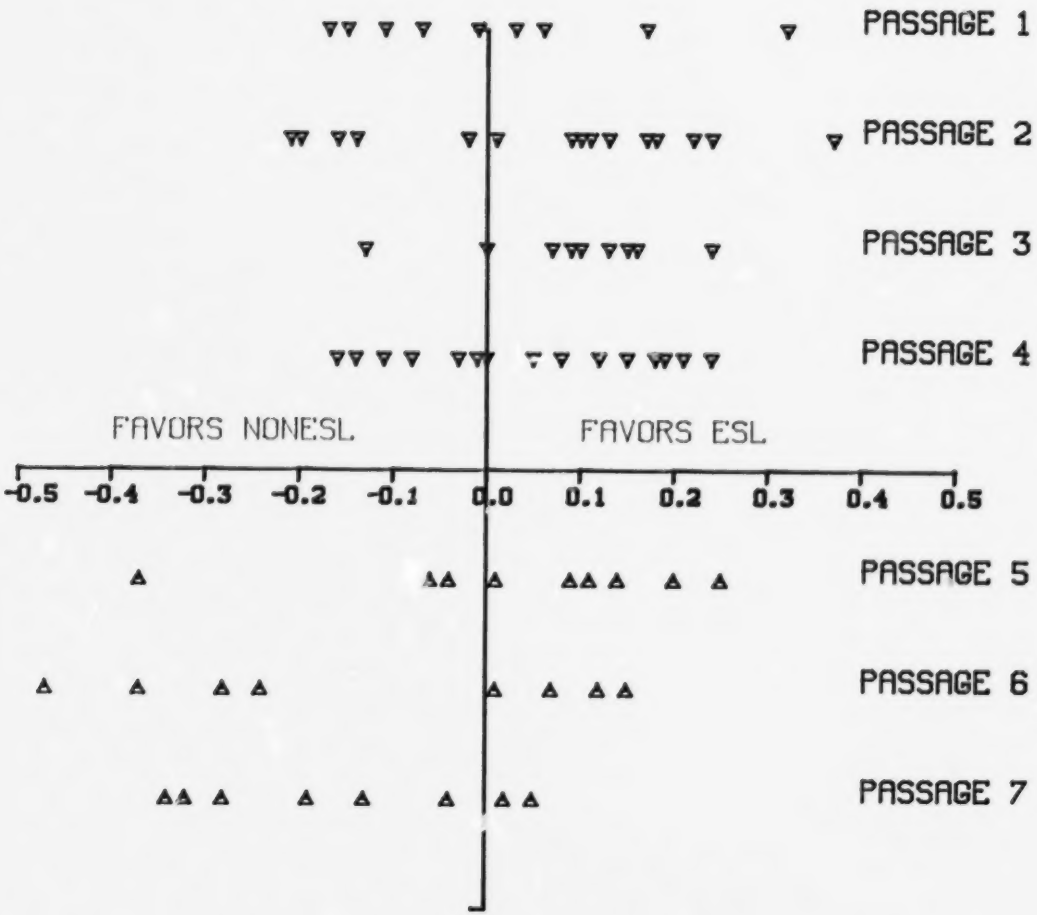

CUMULATIVE FREQUENCY POLYGONS OF ENGLISH USAGE
SCORES FROM ESL AND NONESL STUDENTS
CUMULATIVE PERCENT
RAW SCORE
ESL
NON-ESL
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
72

# FIGURE 2

CUMULATIVE FREQUENCY POLYGONS OF MATH USAGE SCORES FROM ESL AND NONESL STUDENTS



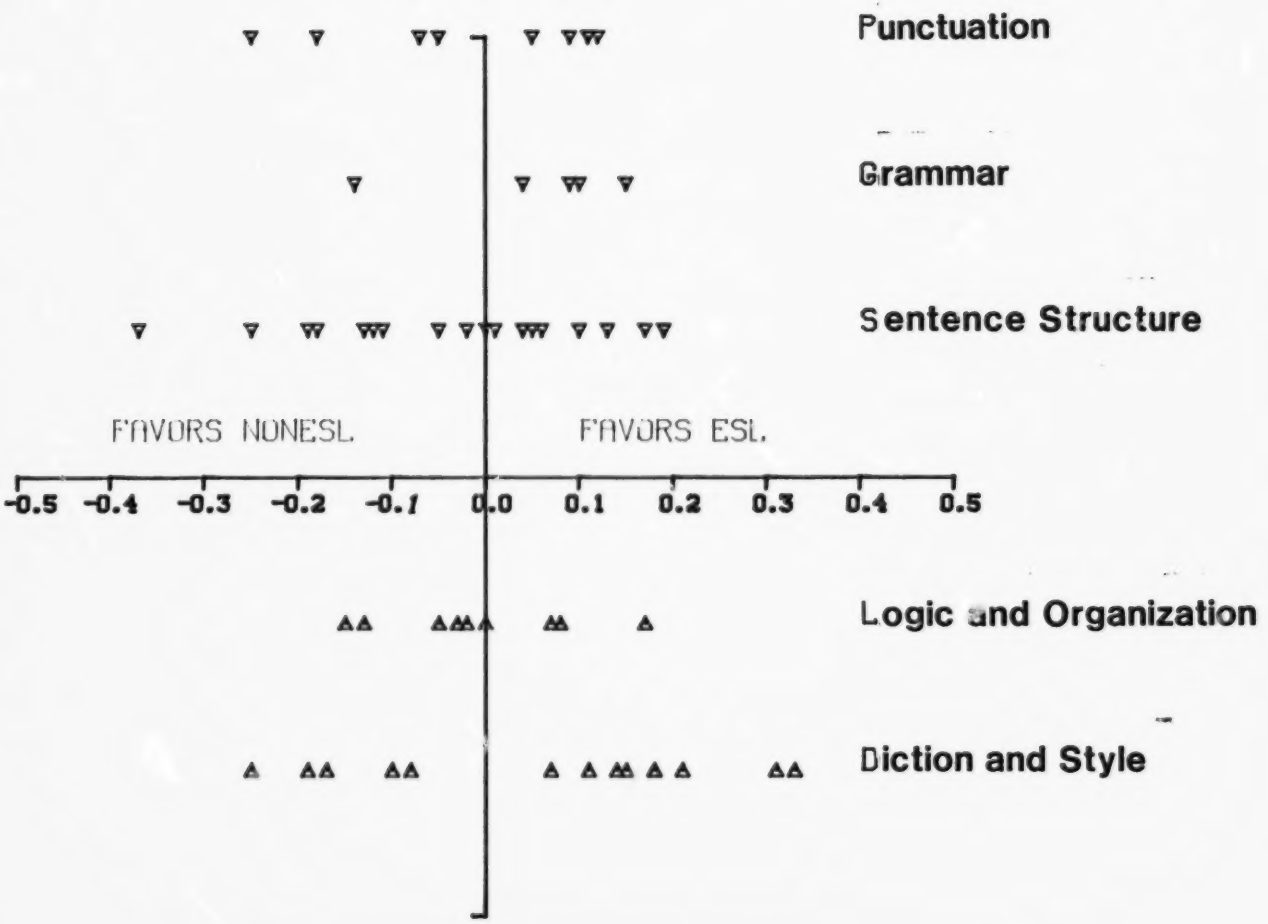
# FIGURE 3




MANTEL-HAENSZEL Z PASSAGE SUMMARY

AAP ENGLISH 28A OCT 1986

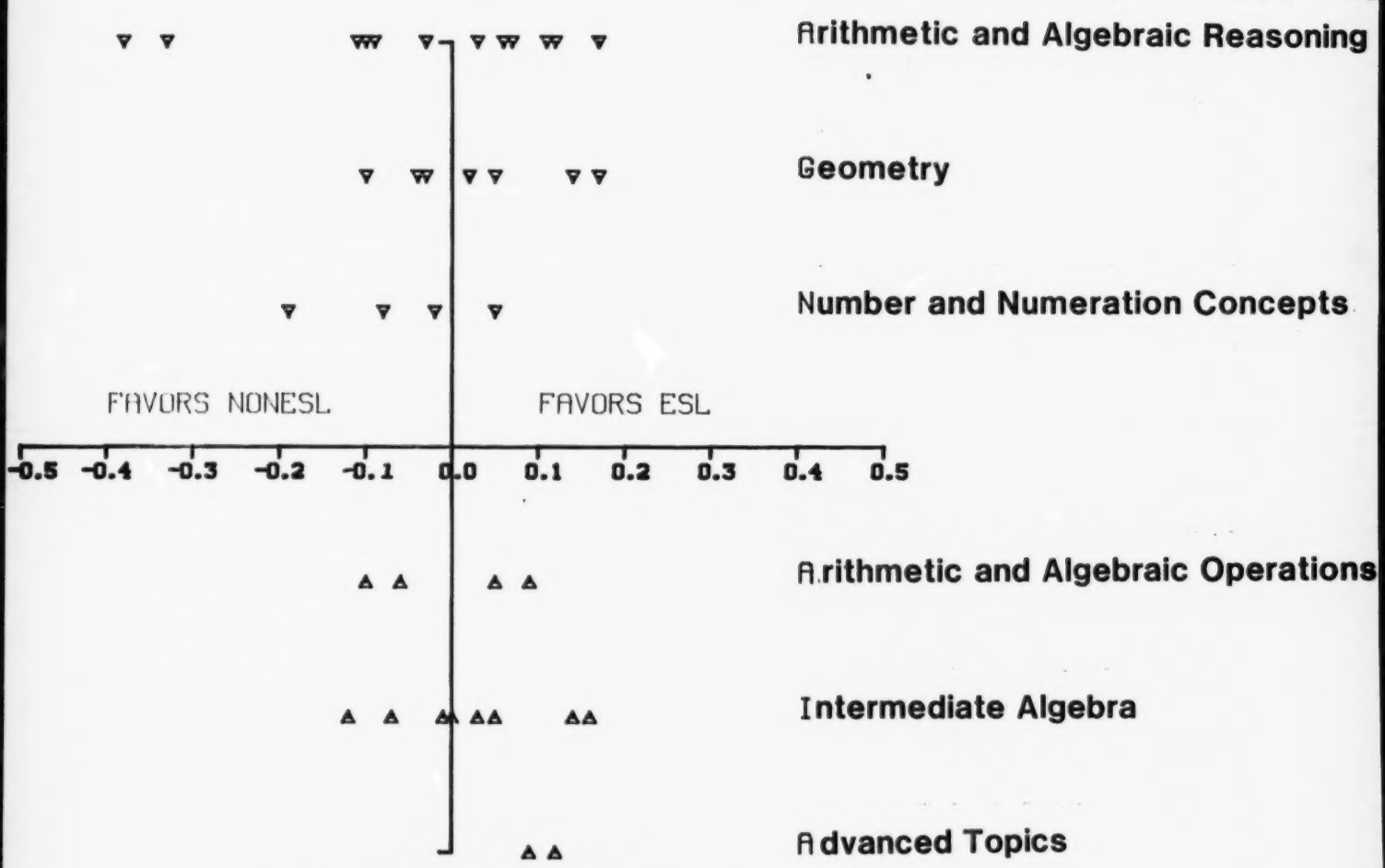
# FIGURE 4




MANTEL-HAENSZEL Z CONTENT SUMMARY

AAP ENGLISH 28A OCT 1986

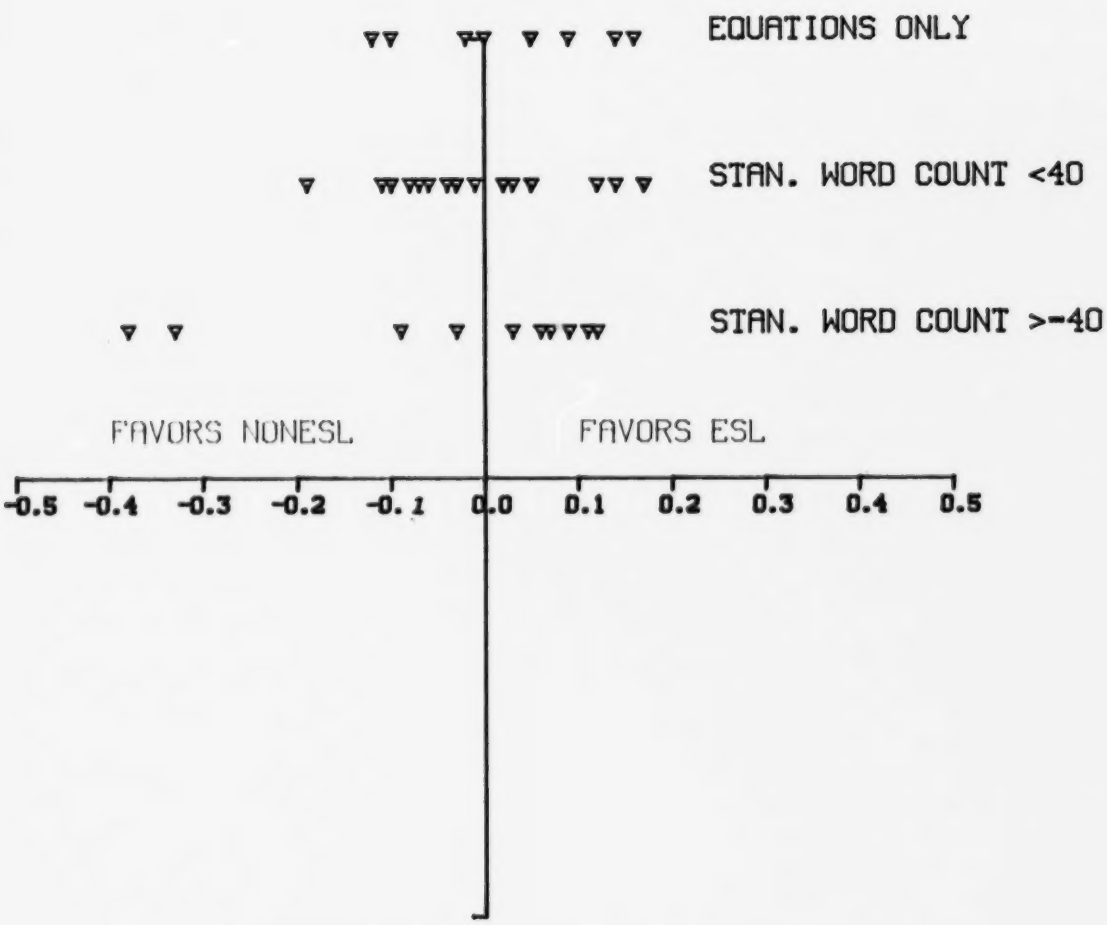
# FIGURE 5




MANTEL-HAENSZEL Z CONTENT SUMMARY

AAP MATH 28A OCT 1986

# FIGURE 6




MANTEL-HAENSZEL Z WORD COUNT SUMMARY

AAP MATH 28A OCT 1986